# **Overall discovery process**

**CIT discovery set** 

537 Afffymetrix U133Plus2.0 transcriptomes



## **Quality control**

AffyQCReport (Rpack bioconductor) Normalization Probeset filtering

no sample removed

52188 probesets



# **Subgroup discovery**

1- Unsupervised probeset selection 244 probests (188 genes)

>>coefficient of variation

2- preliminary clustering and sample coreset

>> convergence of 3 classification methods

>> hierarchical, kmeans (Rpack stat), Gaussian mixture models (Rpack mclust)

3- Identification of a molecular signature

>> analysis of variance

4- Final clustering and sample coreset definition

>> repeat of step 2

394 samples first selection

375 probesets (256 genes)

355 samples coreset selection



#### Molecular classification of breast cancers

>> definition of 6 molecular subgroups



## Classification of all series

definition of a classification rule based on the 256 genes and the 355 tumor coreset



## Association with molecular and bioclinical features

ER/PR/ERBB2 (IHC and RNA expression)

TP53 mutation

CNA profile: 488 tumors analyzed by array-CGH

SBR grading, histotype, age, outcome, site of metastasis